

PCT

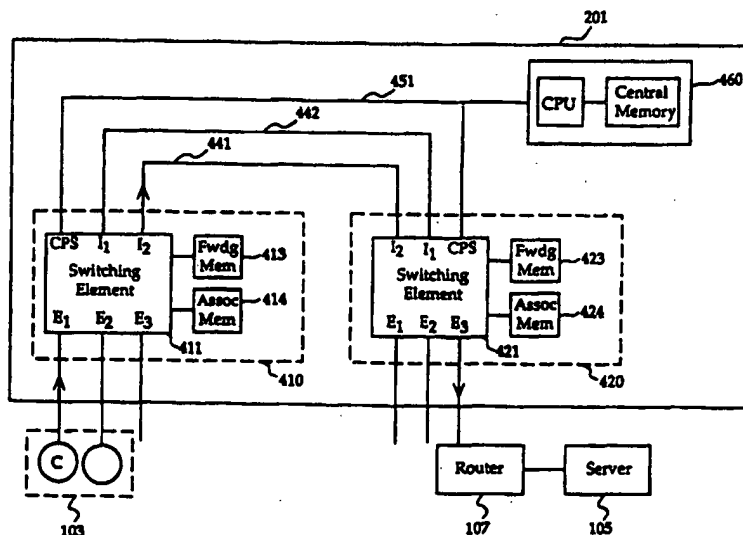
WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : H04L 12/28, 12/56, H04Q 5/22		A1	(11) International Publication Number: WO 99/00938
			(43) International Publication Date: 7 January 1999 (07.01.99)
(21) International Application Number: PCT/US98/13205		(81) Designated States: JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).	
(22) International Filing Date: 24 June 1998 (24.06.98)			
(30) Priority Data: 08/885,114 30 June 1997 (30.06.97) US		Published With international search report.	
(71) Applicant: SUN MICROSYSTEMS, INC. [US/US]; 901 San Antonio Road, Palo Alto, CA 94303 (US).			
(72) Inventors: HENDEL, Ariel; 7537 Newcastle Drive, Cupertino, CA 95014 (US). MULLER, Shimon; 983 La Mesa Tr., Cupertino, CA 94086 (US). ZAUMEN, William, T.; 912 Clara Drive, Palo Alto, CA 94303 (US). YEUNG, Louise; 110 Rogers Avenue, San Carlos, CA 94070 (US).			
(74) Agents: HYMAN, Eric, S. et al.; Blakely, Sokoloff, Taylor & Zafman, 7th floor, 12400 Wilshire Boulevard, Los Angeles, CA 90025-1026 (US).			

(54) Title: ROUTING IN A MULTI-LAYER DISTRIBUTED NETWORK ELEMENT



(57) Abstract

A multi-layer distributed network element (201) for relaying packets according to known routing protocols. A distributed architecture of multiple subsystems (210) delivers routing at wire-speed performance across subnetworks. Each subsystem (210) includes a forwarding memory (213) and an associated memory (214) and is configured to identify unicast and multicast packets for routing purposes, modify the packets in hardware, including replace VLAN information, and forward the packets to the next hop. The routing decisions are made in the inbound subsystem, and packets are forwarded, if necessary given the network topology, through a separate outbound subsystem.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

BACKGROUND

1. Field of the Invention

This invention relates generally to communication systems that couple computers, and more specifically to relaying messages through a network
5 element.

2. Description of Related Art

Communication between computers has become an important aspect of everyday life in both private and business environments. Computers converse with each other based upon a physical medium for transmitting the
10 messages back and forth, and upon a set of rules implemented by electronic hardware attached to and programs running on the computers. These rules, often called protocols, define the orderly transmission and receipt of messages in a network of connected computers.

A local area network (LAN) is the most basic and simplest network that
15 allows communication between a source computer and destination computer. The LAN can be envisioned as a cloud to which computers (also called endstations or end-nodes) that wish to communicate with one another are attached. At least one network element will connect with all of the endstations in the LAN. An example of a simple network element is the
20 repeater which is a physical layer relay that forwards bits. The repeater may have a number of ports, each endstation being attached to one port. The repeater receives bits that may form a packet of data that contains a message from a source endstation, and blindly forwards the packet bit-by-bit. The bits

are then received by all other endstations in the LAN, including the destination.

A single LAN, however, may be insufficient to meet the requirements of an organization that has many endstations, because of the limited number of physical connections available to and the limited message handling capability of a single repeater. Thus, because of these physical limitations, the repeater-based approach can support only a limited number of endstations over a limited geographical area.

The capability of computer networks, however, has been extended by connecting different subnetworks to form larger networks that contain thousands of endstations communicating with each other. These LANs can in turn be connected to each other to create even larger enterprise networks, including wide area network (WAN) links.

To facilitate communication between subnetworks in a larger network, more complex electronic hardware and software have been proposed and are currently used in conventional networks. Also, new sets of rules for reliable and orderly communication among those endstations have been defined by various standards based on the principle that the endstations interconnected by suitable network elements define a network hierarchy, where endstations within the same subnetwork have a common classification. A network is thus said to have a topology which defines the features and hierarchical position of nodes and endstations within the network.

The interconnection of endstations through packet switched networks has traditionally followed a peer-to-peer layered architectural abstraction. In such a model, a given layer in a source computer communicates with the

same layer of a peer endstation (usually the destination) across the network. By attaching a header to the data unit received from a higher layer, a layer provides services to enable the operation of the layer above it. A received packet will typically have several headers that were added to the original
5 payload by the different layers operating at the source.

There are several layer partitioning schemes in the prior art, such as the Arpanet and the Open Systems Interconnect (OSI) models. The seven layer OSI model used here to describe the invention is a convenient model for mapping the functionality and detailed implementations of other models.
10 Aspects of the Arpanet, however, (now redefined by the Internet Engineering Task Force, or IETF) will also be used in specific implementations of the invention to be discussed below.

The relevant layers for background purposes here are Layer 1 (physical), Layer 2 (data link), and Layer 3 (network), and to a limited extent Layer 4
15 (transport). A brief summary of the functions associated with these layers follows.

The physical layer transmits unstructured bits of information across a communication link. The repeater is an example of a network element that operates in this layer. The physical layer concerns itself with such issues as
20 the size and shape of connectors, conversion of bits to electrical signals, and bit-level synchronization.

Layer 2 provides for transmission of frames of data and error detection. More importantly, the data link layer as referred to in this invention is typically designed to "bridge," or carry a packet of information across a single
25 hop, i.e., a hop being the journey taken by a packet in going from one node to

another. By spending only minimal time processing a received packet before sending the packet to its next destination, the data link layer can forward a packet much faster than the layers above it, which are discussed next. The data link layer provides addressing that may be used to identify a source and a destination between any computers interconnected at or below the data link layer. Examples of Layer 2 bridging protocols include those defined in IEEE 802 such as CSMA/CD, token bus, and token ring (including Fiber Distributed Data Interface, or FDDI).

Similar to Layer 2, Layer 3 also includes the ability to provide addresses of computers that communicate with each other. The network layer, however, also works with topological information about the network hierarchy. The network layer may also be configured to "route" a packet from the source to a destination using the shortest path. Finally, the network layer can control congestion by simply dropping selected packets, which the source might recognize as a request to reduce the packet rate.

Finally, Layer 4, the transport layer, provides an application program such as an electronic mail program with a "port address" which the application can use to interface with Layer 3. A key difference between the transport layer and the lower layers is that a program on the source computer carries a conversation with a similar program on the destination computer, whereas in the lower layers, the protocols are between each computer and its immediate neighbors in the network, where the ultimate source and destination endstations may be separated by a number of intermediate nodes. Examples of Layer 4 and Layer 3 protocols include the Internet suite of protocols such as TCP (Transmission Control Protocol) and IP (Internet Protocol).

Endstations are the source and ultimate destination of a packet, whereas a node refers to an intermediate point between the endstations. A node will typically include a network element which has the capability to receive and forward messages on a packet-by-packet basis.

5 Generally speaking, the larger and more complex networks typically rely on nodes that have higher layer (Layers 3 and 4) functionalities. A very large network consisting of several smaller subnetworks must typically use a Layer 3 network element known as a router which has knowledge of the topology of the subnetworks.

10 A router can form and store a topological map of the network around it based upon exchanging information with its neighbors. If a LAN is designed with Layer 3 addressing capability, then routers can be used to forward packets between LANs by taking advantage of the hierarchical routing information available from the endstations. Once a table of endstation addresses and
15 routes has been compiled by the router, packets received by the router can be forwarded after comparing the packet's Layer 3 destination address to an existing and matching entry in the memory.

20 The router operates by parsing the header of a received packet, making decisions based on a routing table inside the router, and forwarding the packet, with any required header modifications, to the next node or endstation. Thus, the packet will go through several such "hops" before reaching its destination where a hop is defined as the packet traveling from one node or endstation to another node or endstation.

25 In comparison to routers, bridges are network elements operating in the data link layer (Layer 2) rather than Layer 3. They have the ability to

forward a packet based only on the Layer 2 address of the packet's destination, typically called the medium access control (MAC) address. Generally speaking, bridges do not modify the packets. Bridges forward packets in a flat network having no hierarchy without any cooperation from the endstations.

- 5 Hybrid forms of network elements also exist, such as routers and switches. A router is a router which can also perform as a bridge. The term switch refers to a network element which is capable of forwarding packets at high speed with functions implemented in hardwired logic as opposed to a general purpose processor executing instructions. Switches come in many
10 flavors, operating at both Layer 2 and Layer 3.

- Having discussed the current technology of networking in general, the limitations of such conventional techniques will now be addressed. With an increasing number of users requiring increased bandwidth from existing networks due to multimedia applications to run on the modern day Internet,
15 modern and future networks must be able to support a very high bandwidth and a large number of users. Furthermore, such networks should be able to support multiple traffic types such as voice and video which typically require different service characteristics. Statistical studies show that the network domain, i.e., a group of interconnected LANs, as well as the number of
20 individual endstations connected to each LAN, will grow at a faster rate in the future. Thus, more network bandwidth and more efficient use of resources is needed to meet these requirements.

- Building networks using Layer 2 elements such as bridges provides fast packet forwarding between LANs but has no flexibility in traffic isolation,
25 redundant topologies, and end-to-end policies for queuing and access control.

Endstations in a subnetwork can invoke conversations based on either Layer 3 or Layer 2 addressing. As bridges forward packets based on only Layer 2 parsing, they provide simple yet speedy forwarding services. However, the bridge does not support the use of high layer handling directives including queuing, priority, and forwarding constraints between endstations in the same subnetwork.

A prior art solution to enhancing bridge-like conversations within a subnetwork relies on a network element that uses a combination of Layer 2 and upper layer headers. In that system, the Layer 3 and Layer 4 information of an initial packet are examined, and a "flow" of packets is predicted and identified using a new Layer 2 entry in the forwarding memory, with a fixed quality of service (QOS). Thereafter, subsequent packets are forwarded at Layer 2 speed (with the fixed QOS) based upon a match of the Layer 2 header with the Layer 2 entry in the forwarding memory. Thus, no entries with Layer 3 and Layer 4 headers are placed in the forwarding memory to identify the flow.

However, consider the scenario where there are two or more programs communicating between the same pair of endstations, such as an electronic mail program and a video conferencing session. If the programs have dissimilar QOS needs, the prior art scheme just presented will not support different QOS characteristics between the same pair of endstations, because the prior art scheme does not consider information in Layer 3 and Layer 4 when forwarding. Thus, there is a need for a network element that is flexible enough to support independent priority requests from applications running on endstations connected to the same subnetwork.

The latter attributes may be met using Layer 3 elements such as routers. But packet forwarding speed is sacrificed in return for the greater intelligence and decision making capability provided by the router. Therefore, networks are often built using a combination of Layer 2 and Layer 3 elements.

- 5 The role of the server has multiplied with browser-based applications that use the Internet, thus leading to increasing variation in traffic distribution. When the role of the server was narrowly limited to a file server, for example, the network was designed with the client and the file server in the same subnetwork to avoid router bottlenecks. However, more
10 specialized servers like World Wide Web and video servers are typically not on the client's subnetwork, such that crossing routers is unavoidable. Therefore, the need for packets to traverse routers at higher speeds is crucial. The choice of bridge versus router typically results in a significant trade-off, lower functionality when using bridges, and lower speed when using routers.
15 Furthermore, the service characteristics within a network are no longer homogenous, as the performance of a server becomes location dependent if its traffic patterns involve routers.

- Therefore, there is a need for a network element that can handle changing network conditions such as topology and message traffic yet make
20 efficient use of high performance hardware to switch packets based on their Layer 2, Layer 3, and Layer 4 headers. The network element should be able to operate at bridge-like speeds, yet be capable of routing packets across different subnetworks and provide upper layer functionalities such as quality of service.

SUMMARY

The invention is an apparatus and related method for relaying packets by a multi-layer distributed network element according to known routing protocols.

5 The invention is directed at a multi-layer distributed network element (MLDNE) for receiving and forwarding packets using known routing protocols. The MLDNE has a number of subsystems that are coupled by internal links. Each subsystem has a forwarding memory and associated memory. The memories associate packet header information including
10 addresses with routing information. A subsystem also includes external ports that connect with neighboring nodes and endstations, and internal ports that connect with other subsystems through the internal links.

When a packet is received by a first "inbound" subsystem, the subsystem determines whether the packet should be routed based upon a first
15 header portion, including a Layer 2 destination address of the received packet, matching a Layer 2 address of the MLDNE. If the first header portion of the received packet matches the MLDNE address, then the first subsystem determines, using its forwarding memory, whether a route has been previously determined for a second header portion, including Layer 3 source
20 and destination addresses, of the received packet.

If a type 2 entry in the forwarding memory matches the received packet's second header portion, then a neighbor node's Layer 2 address (found in associated memory) replaces the Layer 2 destination address of the packet. The neighbor node's address was previously stored in the associated memory

as part of the routing information associated with the matching type 2 entry. In addition to Quality of Service information, the routing information in the associated memory also identifies the external ports of the inbound subsystem that connect with the neighbor node. If the neighbor node is connected to a subsystem other than the inbound subsystem, the situation would have been recognized at the time the matching type 2 entry was created such that the associated memory would identify the internal port of the inbound subsystem, rather than external port, that connects with the other subsystem to which the neighbor node or endstation is connected.

10 When the packet is received over the internal link by a second subsystem, the packet is forwarded to the neighbor node in response to the packet's new first header portion matching a type 1 entry in the second forwarding memory. The type 1 entry in the second subsystem contains the address of the neighbor node or endstation and had been created
15 independently of the matching type 2 entry of the inbound subsystem.

After determining that a received packet should be routed, the inbound subsystem also generates a first control signal which indicates to the external port that eventually forwards the packet that a third header portion identifying the packet's source be modified before sending the packet to the
20 neighbor node. A Layer 2 source address of the packet is replaced with a source address associated with the external port. The control signal is also passed over an internal link to the second subsystem if the neighbor node is reachable through that subsystem.

The invention's distributed architecture can also be configured to
25 support routing of multicast packets. Once a multicast routable packet has

been identified in the inbound subsystem, a second control signal may be sent across an internal link in response to which the second subsystem performs a type 2 search of the forwarding memory (based on the network layer and higher layer headers of the packet). If a matching type 2 entry is found, then

5 the external ports of the second subsystem check the first control signal (also received from the inbound subsystem) to see if the source address of the packet needs to be replaced, and the packet is then forwarded with the appropriate modifications to its headers. The first control signal may also be received and checked by the external ports of the inbound system where the

10 multicast destination group includes nodes/endstations connected to the inbound subsystem.

The invention's search engine, forwarding engine, and data structures are organized in a way that supports bridging and routing functions simultaneously, where if routing criteria are not met for a received packet,

15 then bridging functions are provided automatically.

In its present embodiment, the invention is implemented with the data link layer (Layer 2), the network layer (Layer 3) and higher layers including the transport layer (Layer 4).

DRAWINGS

The foregoing aspects and other features of the invention will be better understood by referring to the figures, detailed description, and claims below where:

5 **Figure 1** is a high level view of an exemplary network application of a multi-layer distributed network element (MLDNE) of the invention.

Figure 2 is an internal view of the MLDNE as an embodiment of the invention.

10 **Figure 3** illustrates an exemplary forwarding and associated memory of a subsystem in the MLDNE, including associated data for the routing of packets, according to another embodiment of the invention.

Figure 4 is a block diagram of an embodiment of the MLDNE having only two subsystems and acting as a router between a client and a server.

15 **Figure 5** is a flow diagram of processing a received packet for routing purposes by the invention's network element.

Figure 6 is a continuation of the flow diagram in **Figure 5** and includes steps performed in processing a unicast packet.

Figure 7 shows exemplary steps and operations performed by the invention's network element for routing a multicast packet.

20 **Figure 8A** is a simplified block diagram of a packet structure utilized in one embodiment of the invention.

Figure 8B is a structure for header field replacement of packets by the invention.

DETAILED DESCRIPTION

As shown in the drawings by way of illustration, the invention defines a network element that is used to interconnect a number of nodes and endstations in a variety of different ways. In particular, an application of the multi-layer distributed network element (MLDNE) would be to route packets according to predefined routing protocols over a homogenous data link layer such as the IEEE 802.3 standard, or Ethernet. Figure 1 illustrates the invention's use as a router in a network where the MLDNE 201 couples a client C to the Router 107 which in turn couples with the Server 105. The MLDNE 201 can interconnect a number of desktop units (endstations), while acting as an intermediate node, through its external connections 217. The MLDNE 201 is capable of providing a high performance communication path between servers and desktop units while acting as a router, where the Server 105 and the client C reside in different LANs.

The MLDNE's distributed architecture can be configured to route message traffic in accordance with a number of known routing algorithms such as RIP and OSPF. In a preferred embodiment, the MLDNE is configured to handle message traffic using the Internet suite of protocols, and more specifically the Transmission Control Protocol (TCP) and the Internet Protocol (IP) over the Ethernet LAN standard and medium access control (MAC) data link layer. The TCP is also referred to here as an exemplary Layer 4 protocol, while the IP is referred to repeatedly as a Layer 3 protocol. However, other protocols can be used to implement the concepts of the invention.

In a first embodiment of the invention's MLDNE, a network element is configured to implement packet routing functions in a distributed manner,

i.e., different parts of a function are performed by identical building block subsystems in the MLDNE, while the final result of the functions remains transparent to the external nodes and endstations. As will be appreciated from the discussion below and the diagram in Figure 2, the MLDNE has a
5 scalable architecture which allows the designer to increase the number of external connections by adding additional subsystems.

As illustrated in block diagram form in Figure 2, the MLDNE 201 contains a number of identical subsystems 210 that are fully meshed and interconnected using a number of internal links 241 to create a larger network
10 element. At least one internal link couples any two subsystems. Each subsystem 210 includes a forwarding memory 213 and an associated memory 214. The forwarding memory 213 stores an address table used for matching with the headers of received packets. The associated memory stores data associated with each entry in the forwarding memory that is used to identify
15 forwarding attributes for forwarding the packets through the MLDNE. A number of external ports (not shown) having input and output capability interface the external connections 217. Internal ports (not shown) also having input and output capability in each subsystem couple the internal links 241. In the preferred embodiment, the external and internal ports lie within a
20 hardwired-logic switching element 211 implemented by an application specific integrated circuit (ASIC).

A received packet arrives at an inbound subsystem through one of the external connections 217, and will be forwarded to a node or endstation outside the MLDNE through another external connection in an outbound
25 subsystem. The outbound and inbound subsystems can be either the same or different subsystems.

Referring to Figure 2, the MLDNE 201 includes a central processing system (CPS) 260 that is coupled to the individual subsystems 210 through a communication bus 251 such as the Peripheral Components Interconnect (PCI). The CPS 260 includes a central processing unit (CPU) 261 coupled to a central memory 263. Central memory 263 includes a copy of the entries contained in the individual forwarding memories 213 of the various subsystems. The CPS has a direct control and communication interface to each subsystem 210. The CPS is also configured with a number of routing protocols that are used to identify a neighbor node as part of a route for forwarding a received packet to its ultimate destination, normally specified in the Layer 3 destination address of the packet. Other responsibilities of the CPS 260 include setting data path resources such as packet buffers between the different subsystems. Finally, the CPS 260 performs the important task of determining whether or not a type 2 entry should be added to the forwarding memory of each individual subsystem.

Figure 3 takes a closer look at the forwarding and associated memories in each subsystem. The forwarding memory includes a number of entries of two types, type 2 entry 321 and type 1 entry 301. Each entry in the forwarding memory includes data to be compared with the headers of received packets. For the particular embodiment of TCP/IP, the data fields for each type 2 entry 321 include a class field 323, an IP source field 325, an IP destination field 327, an application source port 333, an application destination port 335, and an Inbound Port field 337. For the type 1 entry 301, a class field, a Layer 2 address field, and a VLAN identification (VID) field are shown in the exemplary embodiment. Of course, additional header information and similar definitions using alternate network and transport layer protocols can be

developed and included in each entry and used for matching the headers of received packets, as will be apparent to one skilled in the art.

Associated with each type 2 entry 321 and type 1 entry 301 are associated data stored in associated memory 214. The associated data fields contain
5 information needed to forward a matching packet received by the subsystem. The subsystem port field 347 identifies the internal or external ports of the subsystem used for forwarding the matching packet to the neighboring node in the next hop. The next hop address field 357 identifies the neighbor node's Layer 2 address which replaces the original Layer 2 destination address of a
10 received unicast packet to be routed. A priority field 345 is used for queuing purposes by the external port which actually sends the packet outside the MLDNE. The age fields 343 and 344 help minimize the number of entries in the forwarding memory by indicating that a recently received packet has matched the corresponding type 1 or type 2 entry.

15 A NEW VID address field 353 allows the MLDNE to be configured to support virtual LANs (VLANs). The associated data also includes a NEW VLAN identification (VID) TAG field, used to notify the subsystem of a need to change the packet's VID, particularly when forwarding the packet across subnetworks. The inbound subsystem in response will either insert a new
20 tag, or replace an existing tag with the value in the NEW VID field. For example, when routing between VLANs requires the forwarded packet's tag to be different from the received packet's tag, then the NEW VID field will contain the replacement tag for the subsystem to replace before forwarding the packet.

Whenever a packet is sent across an internal link, additional control information may be made available over the internal link to the outbound subsystem receiving the packet. Such information, in addition to the sa_replace bit discussed below, includes an orig_tag bit which indicates whether or not the received packet was originally tagged with VLAN information, a mod_tag bit which indicates whether the tag was modified by the inbound subsystem, and a dont_tag bit which indicates that the received packet should not be tagged by the outbound subsystem.

Finally, the associated memory can be configured to include a multicast route field 355 which activates multicast routing capability in the subsystem as further explained below.

The routing operation of the MLDNE 201 will be described for an exemplary embodiment using the flow diagram of Figures 5-7 in conjunction with the exemplary network application in Figure 4. References to fields in the forwarding and associated memories are found in Figure 3. In the example below, the journey of a packet is traced beginning with a client C in subnetwork 103 coupled to an external connection of MLDNE 201. The client C sends a packet to server 105 which is identified in the Layer 3 destination address field of the packet's header. The packet must traverse a router 107 which is assumed to have a Layer 2 address known by the MLDNE 201.

Beginning with block 503 in Figure 5, a packet is received by the MLDNE 201 at external port E₁ of the inbound subsystem 410. The packet includes a message originated from a client C having a Layer 3 address in a logically defined network subnetwork 103. Subsystem 410 is configured to recognize that external ports E₁ and E₂ couple the subnetwork 103.

When the packet is received by switching element 411, operation continues with decision block 507 where first header portion, including the Layer 2 destination address in the present embodiment, of the received packet is compared with a router address of the MLDNE 201. The router address may
5 be a Layer 2 address assigned to external port E₁, or a Layer 2 address assigned to the MLDNE as a whole. Normally, the MLDNE will be configured so that each external port is assigned its own router address.

If the first header portion of the received packet matches the router address, then operation proceeds to block 515 where the packet is declared to
10 be a potential unicast route candidate. If, however, the first header portion does not match the router address, then operation proceeds to block 509 where the packet is declared as not being a unicast routable packet. As will be appreciated below, such a packet can still be a multicast packet having a multicast route available in the MLDNE.

15 For a unicast packet of the route class, block 517 performs a search of the forwarding memory 413 for a matching type 2 entry using "route" as the class field 323.

The search of the forwarding memory in block 517 leads to the decision block 521 where the test is whether a type 2 matching entry exists in the
20 forwarding memory 413. If not, then operation proceeds with block 523 where relevant portions of the received packet headers are sent to the CPS via the CPS port in subsystem 410 and the CPS bus 451.

When the CPS 460 receives the portions of the headers of the "missed" packet from subsystem 410 in block 533, the CPS then examines access policies
25 and class of service policies that have been preconfigured in the CPS, and the

CPS Layer 2 and Layer 3 topology tables. The CPS has the option of denying service to the path requested by the received packet, performing the routing function entirely in its own software, or preparing a type 2 entry in the inbound system's forwarding memory for the route.

5 The routing algorithms of the MLDNE 201 are implemented by the CPS. If a unicast route exists or can be readily computed for the received packet, then the CPS decides in decision block 537 to proceed with block 539 and add a route class type 2 entry 321 to the forwarding memory, and associated data to the associated memory, of the inbound subsystem 410. If
10 the neighbor node connects to an external port of the inbound subsystem 410, as determined by the CPS consulting a Layer 2 table in the central memory, then the external port is identified in the new type 2 entry's associated subsystem port field 347. Similarly, if the neighbor node connects to the subsystem 420, then an internal port I_1 or I_2 is identified.

15 Returning to decision block 521, if the packet matches an existing route class type 2 entry in the forwarding memory 413 of the inbound subsystem 410, then the received packet is forwarded as a unicast packet as illustrated in exemplary form in Figure 6.

 Turning now to Figure 6 and staying in the inbound subsystem, the
20 switching element 411 evaluates whether the unicast packet's time to live has been exceeded. A time to live field is assumed to exist in the received packet's headers. If the packet has been circulating through the network too long as indicated by its time to live field, then the inbound subsystem only sends the received packet to the CPS, and then a time exceeded error message in
25 accordance with, for example, the Internal Control Message Protocol (ICMP)

or as discussed in the Request For Comments (RFC) maintained by the Internet community, is generated by the CPS as in block 609.

If, on the other hand, the packet's time to live (TTL) has not been exceeded, then operation continues with block 619 where the TTL is decremented. This modification to the packet's header will normally require compensating the packet's Layer 3 header check sum as in block 621. In block 611, the switching element 411 replaces the Layer 2 destination address of the received packet with the next hop Layer 2 address found in the associated memory corresponding to the matching type 2 entry determined in block 521 of Figure 5.

If the MLDNE 201 is configured to support VLANs, then decision block 615 determines whether a new VLAN identification tag is required by checking the status of the NEW VID tag field 351.

Whether or not the packet is to be forwarded outside the MLDNE by another subsystem (as indicated by the subsystem port field 347 associated with the matching type 2 entry) a first control signal, such as a sa_replace bit, is prepared in block 621. The sa_replace bit will be handed off to the external and internal ports indicated in the subsystem port field 347, and thus may be transferred over an internal link 441, together with the packet, to the subsystem 420. The first control signal will notify the subsystem (either the inbound one or another subsystem) to replace the Layer 2 source address of the packet with the source address of the external port used for forwarding the packet.

In the example of Figure 4, the packet together with any control information, including first control signal, are processed by internal port I₂ in

switching element 411, and delivered to the internal link 441 to connect with the outbound subsystem 420 in block 627. Alternatively, however, the modified packet and control information stay in the inbound subsystem and are processed by an external port, where operation continues in block 630.

5 In block 627, the packet is received over the internal links in outbound subsystem 420. A type 1 matching cycle then begins and decision block 629 is reached to determine whether a matching type 1 entry exists in the forwarding memory 423. If a type 1 entry exists then operation continues with block 630.

10 The operation from block 630 to block 637 are performed by the "outbound" subsystem where the packet leaves the MLDNE, be it the inbound subsystem 410 or a different subsystem 420. If the sa_replace bit, as checked in decision block 630, is set, then the switching element replaces a third header portion, including at least the Layer 2 source address of the
15 received packet, with the Layer 2 address of the external port E₃ through which the packet must be forwarded. The external port E₃ was identified in the associated data (in associated memory) corresponding to either the matching type 1 entry found in block 629 (the packet came across internal link) or the matching type 2 entry found in block 521 (the packet remained in
20 inbound subsystem).

 The MLDNE can be configured so that each external port is assigned a unique Layer 2 address. Alternatively, a single source address may be assigned to the MLDNE as a whole and shared by all external ports. In either case, following the replacement of the third header portion, the cyclic
25 redundancy code (CRC) of the packet's headers is recomputed in block 635 and

the packet is then forwarded to the neighbor node being the router 107 in Figure 4.

In the above example, the packet's journey has been described originating from the client C and traveling through subsystem 410, internal link 441, and subsystem 420 in MLDNE 201. The packet is then received by router 107 and forwarded according to conventional means to server 105. The above, of course, assumed that a route for the server 105 as a destination through router 107 had been previously obtained by the MLDNE 201 using conventional techniques for determining the routes.

The above also covered the situation where although a unicast packet falls within the route class, no type 2 matching entry existed in the inbound subsystem to be used for routing the packet through the MLDNE. Thus, the decision as to whether or not a received packet will be routed is made in the inbound subsystem, in particular, in decision blocks 507 and 521 of Figure 5. Note also that routing policies as well as class of service queuing have the granularity and flexibility of Layer 3 end-to-end addresses and protocol based classification. These routing policies and class of service queuing are identified in the associated data corresponding to each matching type 2 entry, and may be sent across the internal link to a separate outbound subsystem.

Multicast Routing

Having discussed the unicast routing aspects of the invention, the routing features of the invention for multicast packets are now presented while referring once again to the entries in the forwarding and associated memories of Figure 3 and the flow diagram of Figure 7. Although multicast routing in the invention's MLDNE can be supported by similar hardware

structures that implement unicast routing in the MLDNE, multicast does present significantly different problems to the network element designer. For instance, the routing protocols used to derive the type 2 entries in the forwarding memory include protocols such as MOSPF and DVMRP which are well-known in the art. These multicast routing protocols produce a loop-free distribution tree for the packet's group destination network layer multicast address and a source network layer address for the sender.

The MLDNE has a local multicast forwarding rule which yields a number of external ports (and their corresponding subsystems) for forwarding the packet, as a function of a received multicast packet group destination Layer 3 address, source Layer 3 address, and the inbound subsystem port of arrival. This dependency is reflected in the type 2 entry in the forwarding memory of Figure 3 as the fields 327, 325, and 337, respectively, to be matched with a received packet's headers. The inbound port of arrival field 337 is included to prevent forwarding duplicate packets over alternate paths.

To identify a received packet as a candidate for multicast routing, the MLDNE is configured to identify a multicast packet based on at least two criteria. First, the packet headers must match a given class. Second, the packet's headers must match an existing type 2 entry that refers to a multicast group destination address. The matching type 2 entry for the multicast case may be created as a result of executing a multicast registration protocol such as IGMP.

Figure 7 illustrates an exemplary flow diagram for routing a received multicast packet through the MLDNE 201 of Figure 4. When a packet is received by the subsystem 410 and the packet headers match a certain class

and a type 2 entry 321 which has a multicast route field 355 indicating that the entry is for multicast routing, as in block 703, control is transferred to the decision block 705. If the packet's time to live has not been exceeded, then the routing operation continues in block 709 in the inbound subsystem 410 by
5 decrementing the time to live field in the received packet's header. If the packet's TTL was exceeded, then in block 707 the packet may be flooded, not routed, to its VLAN. A packet's VLAN, in general, defines the Layer 2 topology used for flooding, in other words the broadcast domain.

Proceeding to block 711, the inbound subsystem 410 determines
10 whether a new VLAN tag is required for the received packet, based on the NEW VID tag field 351 in the associated memory. If so, then the VID in the Layer 2 header of the packet is replaced with the destination VID of the next hop, as found in the associated memory, as in block 713. Note that block 713 is performed only if the Layer 3 multicast destination address of the received
15 packet refers to endstations that lie within the same VLAN. Such a determination was made by the CPS when the type 2 entry was created.

Whether or not VLANs are supported by the MLDNE, in block 715 the inbound subsystem 410 prepares to notify the external ports that will forward the packets outside the MLDNE of a need to route the packet by setting the
20 first control signal (sa_replace bit) to indicate to the forwarding external ports that the Layer 2 source address of the packet to be forwarded must be replaced with the source address of the external port. Once the changes have been made to the network layer header, in particular, the portion that includes the time to live (TTL) field, the inbound subsystem compensates the packet's
25 header check sum value in block 717. The inbound subsystem 410 then hands off copies of the packet to the external and internal ports of the inbound

subsystem 410 that are identified in the subsystem ports field 347 of the associated memory as corresponding to the matching type 2 entry, as in block 719.

In the case where a copy of the packet traverses an internal link and
5 arrives at a different subsystem 420 in block 720, operation proceeds with decision block 721 where a second control signal, here called the distributed flow (DF or distrib_flow) bit, may be received by the outbound subsystem 420. If the DF bit is set, then a class filter determines the class of the packet, based upon the packet's headers, and a type 2 search (with the identified class) is
10 conducted in block 722.

The distrib_flow construct allows the CPS to define a type 2 entry in the outbound subsystem 420 corresponding to the matching multicast route entry in the inbound subsystem. This allows different priorities to be assigned by the CPS to the different external ports that will service the multicast route, to
15 further control queuing granularity for packets traversing the MLDNE. A force_be bit (placed by the CPS and obtained after a type 2 search in the outbound subsystem) in the associated data of the matching type 2 entry overrides the priority received over the internal link with the packet, such that the packet will be forced to the lowest priority, thus providing some
20 granularity in queuing at the external ports.

If the distrib_flow bit is not set, then a type 1 search is performed on the forwarding memory 423, and the packet is forwarded or flooded accordingly without the type 2 queuing granularity discussed above.

If a matching type 1 or type 2 entry is found, then the packet is handed
25 off to the external ports identified in the associated memory corresponding to

the matching entry. Thereafter, operation proceeds with block 723. Thus, a multicast route requires two type 2 entries to be created by the CPS where the inbound and outbound subsystems are different.

The operations from block 723 to block 729 are performed by the
5 outbound subsystem, be it the subsystem 410 or subsystem 420. The outbound subsystem in decision block 723 determines whether the sa_replace bit has been set to indicate that the Layer 2 source address of each copy of the packet should be replaced with the Layer 2 address of the corresponding external port used for forwarding the packet outside the MLDNE. If not, then the packet
10 may be forwarded using a Layer 2 search result.

If there is an indication to replace the Layer 2 source address for routing purposes, then in block 725, the outbound subsystem, in particular an external port of the outbound subsystem, replaces the Layer 2 source address of the packet with a Layer 2 address of the external port. Operation then proceeds
15 with block 727 where a CRC is recomputed for the modified Layer 2 header, and the packet is forwarded in block 729.

An innovative structure and method for transmitting the packet and control information across the internal link will now be described with reference to Figures 8A and 8B. Figure 8A is a simplified diagram of the
20 packet structure utilized. More particularly, as the inbound subsystem has determined certain information regarding the packet, e.g., routing, it is advantageous to simply convey this information to the outbound subsystem so that subsequent processing, such as the header field replacement, can easily be performed without reperforming the same steps performed by the inbound
25 subsystem. Furthermore, it is desirable to maintain end-to-end error

robustness. Thus, the inbound subsystem encapsulates the packet 800 with control information 805 and a cycle redundancy code (CRC) 810. The outbound system receives the encapsulated packet, determines frame validity using CRC 810, strips the CRC 810 and removes the control information 805
5 to determine the subsequent processing to be performed to output the packet.

The control information includes information to instruct the outbound subsystem how to update the header information, if needed, before output. In the present embodiment, the control information includes the following:

- 10 ◦ replace_sa - when set, indicates that the source address field of the header is to be replaced with the outbound subsystem's output MAC address;
- orig_tag - when set, indicates that the VLAN tag is the original tag the packet arrived with at the inbound subsystem;
- mod_tag - when set, indicates that the VLAN tag the packet arrived
15 with has been modified;
- dont_tag - when set, indicates that the VLAN tag may not be used regardless of the state of the orig_tag and the mod_tag (in the present embodiment, this is typically used when packets arrive from the CPS 260);
- distributed_flow - when set, indicates whether a Layer 3 or Layer 2
20 search should be conducted initially for the packet;
- priority (2) - indicates the queuing priority level in the subsystem external ports for the particular packet;
- reserved (9)

A simplified block diagram illustrating the process for header field replacement of packets communicated through internal links is illustrated in the diagram of Figure 8B. For purposes of explanation, a number of functional elements not relevant to the process of performing header field replacement are not shown or described. However, it is readily apparent to one skilled in the art that the inbound subsystem includes elements to process the received packet prior to transmission to the outbound system and the outbound system includes elements that perform other function in addition to those described herein.

Referring to Figure 8A, the inbound system 825 receives the packet and accesses the memory containing the database (not shown) to obtain information regarding the packet, e.g., if the packet is to be routed or if VLAN routing is supported. Certain control information is generated and provided to the cascading output process (COP) 835 which prepends the control information to the packet and outputs the packet with the prepended control information to the output interface 840 which generates and appends a CRC to encapsulate the packet for output to the outbound subsystem 830. Preferably the output interface is a media access controller (MAC); however, other interfaces could be used.

The outbound subsystem 830 receives the encapsulated packet at the input interface 845, which is preferably a MAC, performs frame validity checking and strips the CRC. The input interface 845 outputs to the cascading input process (CIP) 850 the packet stripped of the CRC and the CIP 850 removes the control information and forwards the packet, stripped of the encapsulating CRC and control information, to the packet memory 855. The control information is stored in the control field 857 corresponding to the

packet stored in the memory 855. The output port process 860 retrieves the packet and the control information from the packet memory 855 and based upon the control information, selectively performs modifications to the packet and issues control signals to the output interface 865 (i.e., MAC).

- 5 In one embodiment, which occurs when the packet is to be routed, the OPP 860 strips the last 4 bytes of the packet corresponding to the CRC and asserts control signals to the MAC 865 to append a CRC and replace the source address with its own MAC address. For example, the OPP 860 issues a replace_SA signal and clears a no_CRC bit in a control word sent to the MAC
- 10 865. In another embodiment, when VLAN routing is supported, depending upon the state of the control signals, the OPP 860 removes the VLAN tag field in the packet, strips the last 4 bytes of the packet corresponding to the CRC and issues a control signal to the MAC 865 to append a CRC. More particularly, the OPP 860 decodes , orig_tag, mod_tag and dont_tag and a
- 15 fourth indicator, tag_enable. Tag_enable is an internal variable which indicates that the network segment connected to this output port does not support VLAN tagging. This variable is determined by a network management mechanism based on the underlying network topology. The result of the decoding process indicates whether the OPP 860 is to strip the tag
- 20 and whether the MAC 865 is to generate a CRC. The OPP decodes according the following table:

dont_tag	tag_enable	orig_tag	mod_tag	strip_tag	regenerate CRC
1	x	0	x	Y	N
1	x	1	x	Y	N
0	0	0	x	Y	N
0	0	1	x	Y	Y
0	1	0	x	N	Y
0	1	1	0	N	N
0	1	1	1	N	Y

Thus if the tag is to be stripped, the OPP 860 removes the tag, preferably as the tag is transferred to the MAC 86. If no CRC is to be generated, the OPP 860 sends a signal indicating that no CRC is to be generated (e.g., set no_CRC) and the MAC 865 transmits the packet as it is received. If the CRC is to be
5 generated, the last 4 bytes are removed from the packet by the OPP 860 a signal to generate the CRC is sent to the MAC 865, (e.g., clear no-CRC).

The MAC 865, based upon the control signals received from the OPP 860, replaces the source address field with its own MAC address and generates a CRC that is appended to the end of the packet as the packet is output.

10 The encapsulation process can potentially extend the packet by a number of bytes. This can negatively affect the capacity of the link. In order to compensate for this capacity loss and also to allow the reception of frames that may longer than standard protocols define, the protocol parameter (in the present embodiment the Ethernet protocol) are fine tuned to reduce the
15 preamble size by 5 bytes, the interpacket gap by 5 bytes and increase the maximum packet size by 10 bytes.

The embodiments of the routing apparatus and methods in the MLDNE 201 described above for exemplary purposes are, of course, subject to other variations in structure and implementation within the capabilities of
20 one reasonably skilled in the art. Thus, the details above should be interpreted as illustrative and not in a limiting sense.

CLAIMS

What is claimed is:

- 1 1. In a network element for receiving and forwarding packets
2 between nodes, the network element having first and second subsystems
3 coupled by an internal link, the subsystems having first and second
4 forwarding memories, respectively, a method for relaying packets comprising
5 the steps of:
6 receiving a packet by the first sub-system, the packet having a first
7 header portion;
8 determining whether the packet should be forwarded in accordance
9 with a routing protocol in response to parsing the first header portion of the
10 packet;
11 searching the first forwarding memory for a first entry that matches a
12 second header portion of the first packet;
13 replacing part of the first header portion of the packet with a next hop
14 address in response to the second header portion of the packet matching the
15 first entry, the next hop address being associated with the first entry, the next
16 hop address being an address of a neighbor node;
17 sending the packet having the next hop address to the second sub-
18 system through the internal link; and
19 forwarding the packet to the neighbor node.

1 2. A method as in claim 1 wherein
2 the first header portion comprises a Layer 2 destination address, and
3 the second header portion comprises a Layer 3 destination address.

1 3. A method as in claim 1 wherein the step of
2 determining whether the packet should be forwarded includes
3 determining whether the first header portion matches an address assigned to
4 the network element.

1 4. A method as in claim 1 further comprising
2 sending a control signal to the second subsystem over the internal link
3 in response to the second header portion matching the first entry;
4 receiving the packet by the second subsystem; and
5 replacing part of a third header portion of the packet with an address of
6 the second subsystem in response to the second subsystem receiving the
7 control signal.

1 5. A method as in claim 4 wherein the third header portion
2 comprises a Layer 2 source address.

1 6. A method as in claim 1, wherein the step of:
2 sending the packet to the second subsystem includes the step of sending
3 the packet to an internal port of the first subsystem, the internal port coupling

4 the internal link, the internal port being identified as a value associated with
5 the first entry.

1 7. A method as in claim 4 wherein the step of:
2 forwarding the packet includes forwarding the packet with the address
3 of the second subsystem.

1 8. A method as in claim 4 wherein the step of:
2 replacing a third header portion of the first packet includes replacing
3 part of the third header portion with a L2 address of a second external port in
4 the second subsystem.

1 9. A method as in claim 1 further comprising the step of:
2 updating a time to live field of the packet in response to determining
3 by the first subsystem that the packet be forwarded in accordance with a
4 routing protocol; and
5 compensating a header checksum of the packet by the first subsystem.

1 10. A method as in claim 4 further comprising the step of:
2 computing a cyclic redundancy code (CRC) of the packet by the second
3 subsystem after replacing part of the third header portion of the packet.

1 11. A method as in claim 4 further comprising the step of:
2 inserting a Virtual local area network IDentification (VID) of the
3 second subsystem into the packet in response to receiving a NEW TAG
4 notification over the internal link.

1 12. A network element for receiving and forwarding packets
2 between nodes, comprising:
3 first subsystem having a first forwarding memory, the first subsystem
4 configured to determine whether a packet should be routed based upon a part
5 of a first header portion of the packet matching an address of the network
6 element;
7 second subsystem having a second forwarding memory; and
8 an internal link coupling the first and second subsystem, wherein
9 the first subsystem is configured to replace the part of the first
10 header portion with a next hop address in response to matching part of a
11 second header portion of the packet with a first entry in the first forwarding
12 memory, the next hop address being an address of a neighbor node, and
13 wherein
14 the first subsystem is further configured to send the packet with
15 the next hop address to the second subsystem over the internal link, and
16 wherein
17 the second subsystem is configured to forward the received
18 packet to the neighbor node in response to the first header portion matching a
19 second entry in the second forwarding memory.

1 13. A network element according to claim 12 wherein
2 the address of the network element is an L2 address of an external port
3 that receives the packet.

1 14. A network element for receiving and forwarding multicast
2 packets between nodes, comprising:
3 first subsystem having a first forwarding memory, the first subsystem
4 configured to determine whether a multicast packet should be multicast
5 routed based upon a first entry in the forwarding memory matching a header
6 portion of the packet, the first subsystem further including a multicast route
7 indication associated with the first entry;
8 second subsystem having a second forwarding memory, the second
9 forwarding memory including a second entry; and
10 an internal link coupling the first and second subsystems for passing
11 the multicast packet from the first subsystem to the second subsystem,
12 wherein the second subsystem is configured to forward a
13 plurality of packets to a plurality of neighbor nodes in response to receiving
14 the multicast packet over the internal link, the second entry matching the
15 header portion of the packet, and replacing a third header portion of each of
16 the plurality of packets with an address of the second subsystem.

1 15. A network element as in claim 14 wherein
2 the first and second entries each comprises network layer addresses.

1 16. A network element as in claim 14 wherein
2 the third header portion of each of the plurality of packets comprises a
3 data link layer source address.

1 17. A network element as in claim 14 wherein
2 the second subsystem receives a second control signal over the internal
3 link from the first subsystem, in response to which a search of the second
4 forwarding memory is conducted.

1 18. A network element as in claim 17 wherein
2 the search results in the second entry matching the header portion of
3 the packet.

1 19. A network element as in claim 14 wherein
2 the internal link is further configured to pass queuing priority
3 information from the first subsystem to the second subsystem.

1 20. A network element as in claim 17 wherein
2 the search results in a third entry matching the header portion, the
3 second subsystem further including a third control indication associated with
4 the third entry in response to which the second subsystem overrides queuing
5 priority indication received from the first subsystem.

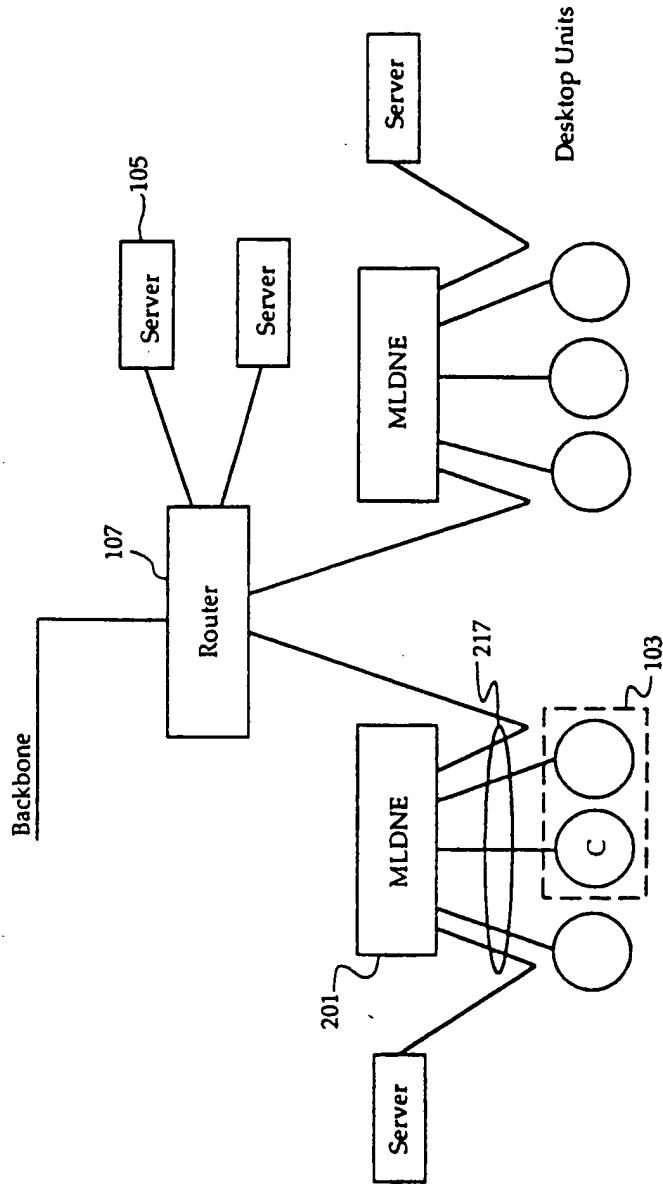
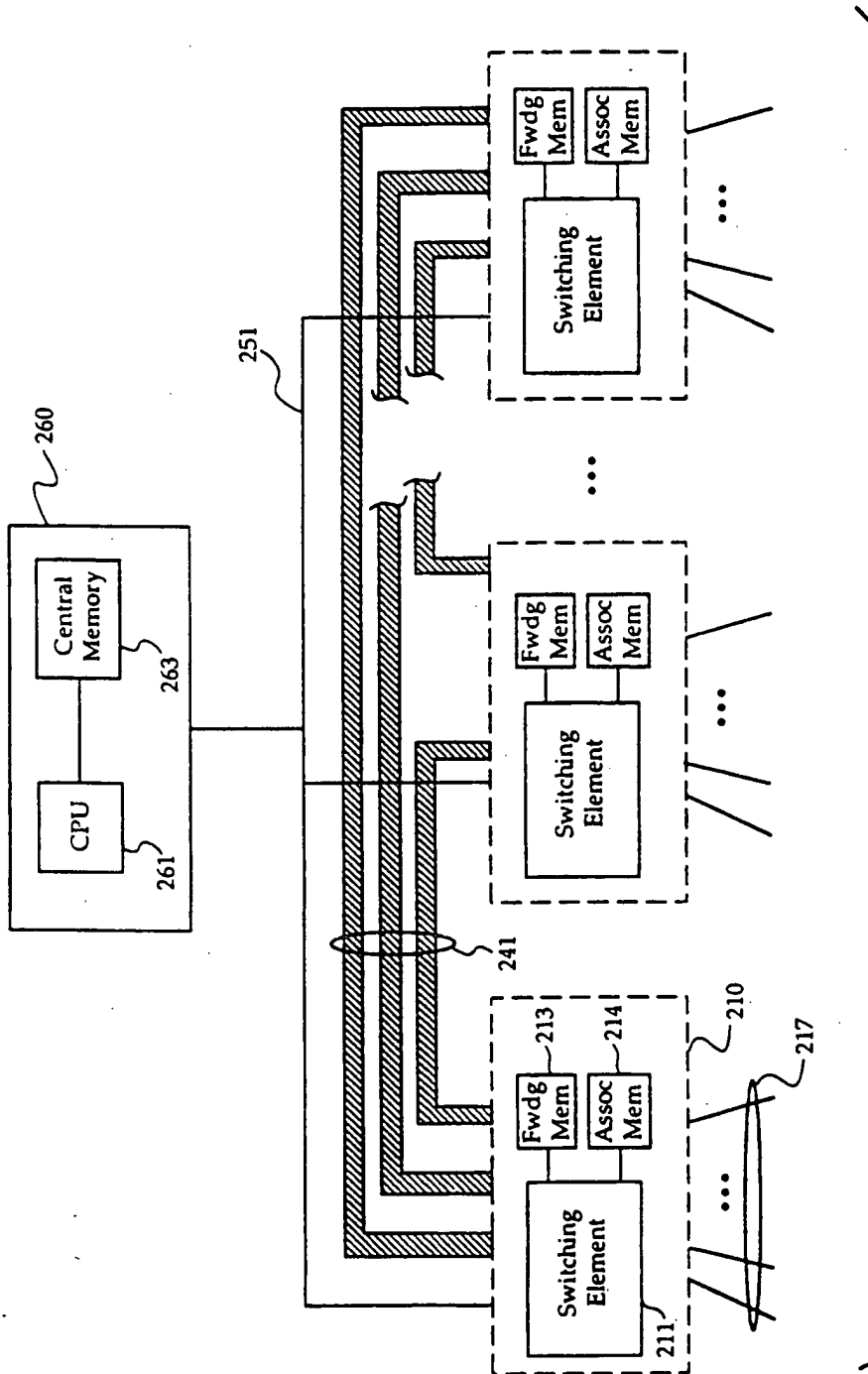


FIGURE 1



To nodes and end-stations

FIGURE 2

201

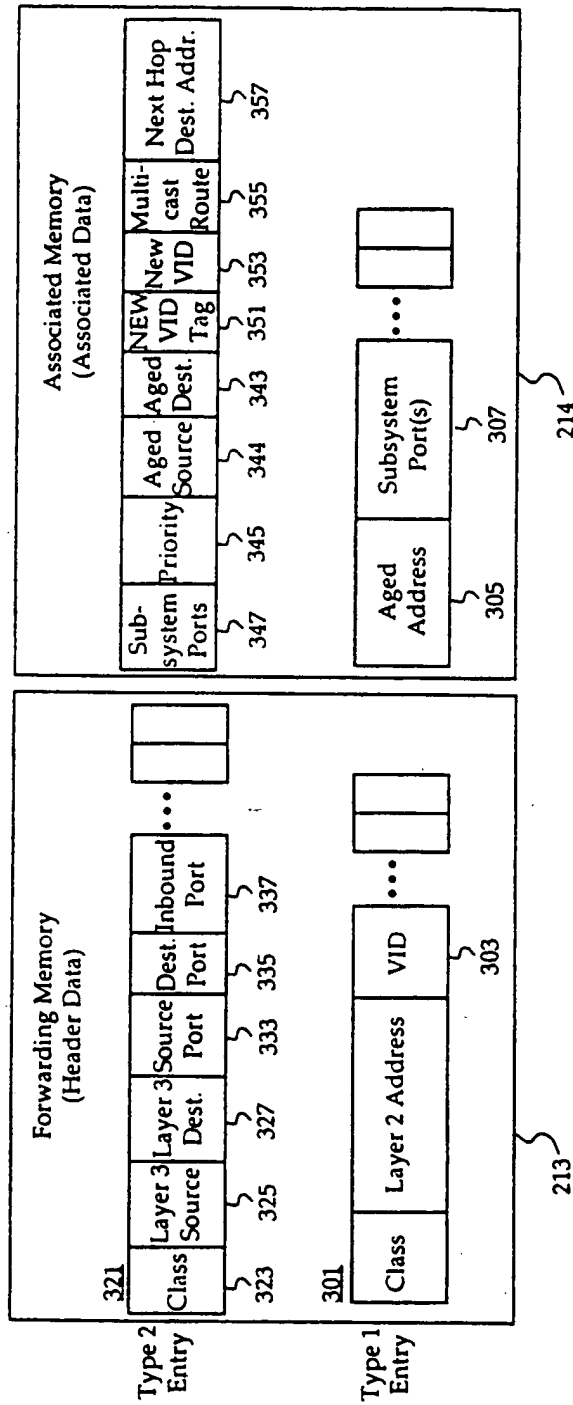


FIGURE 3

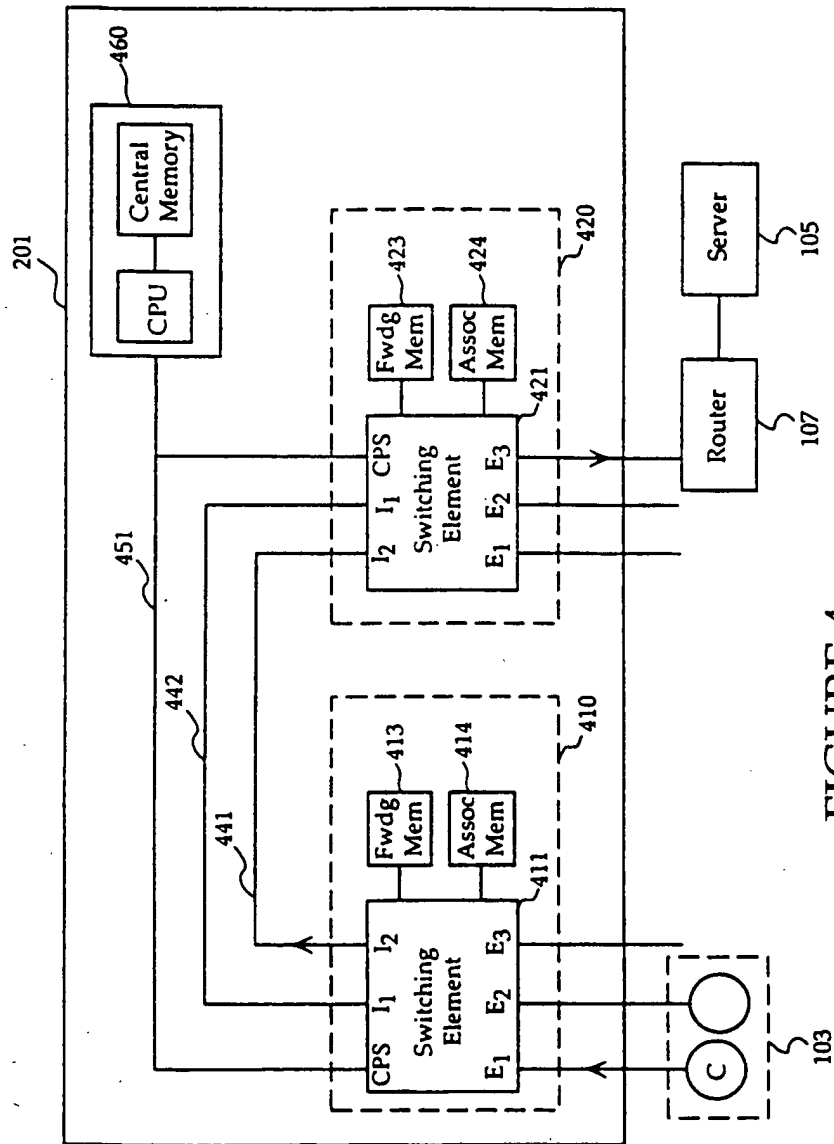


FIGURE 4

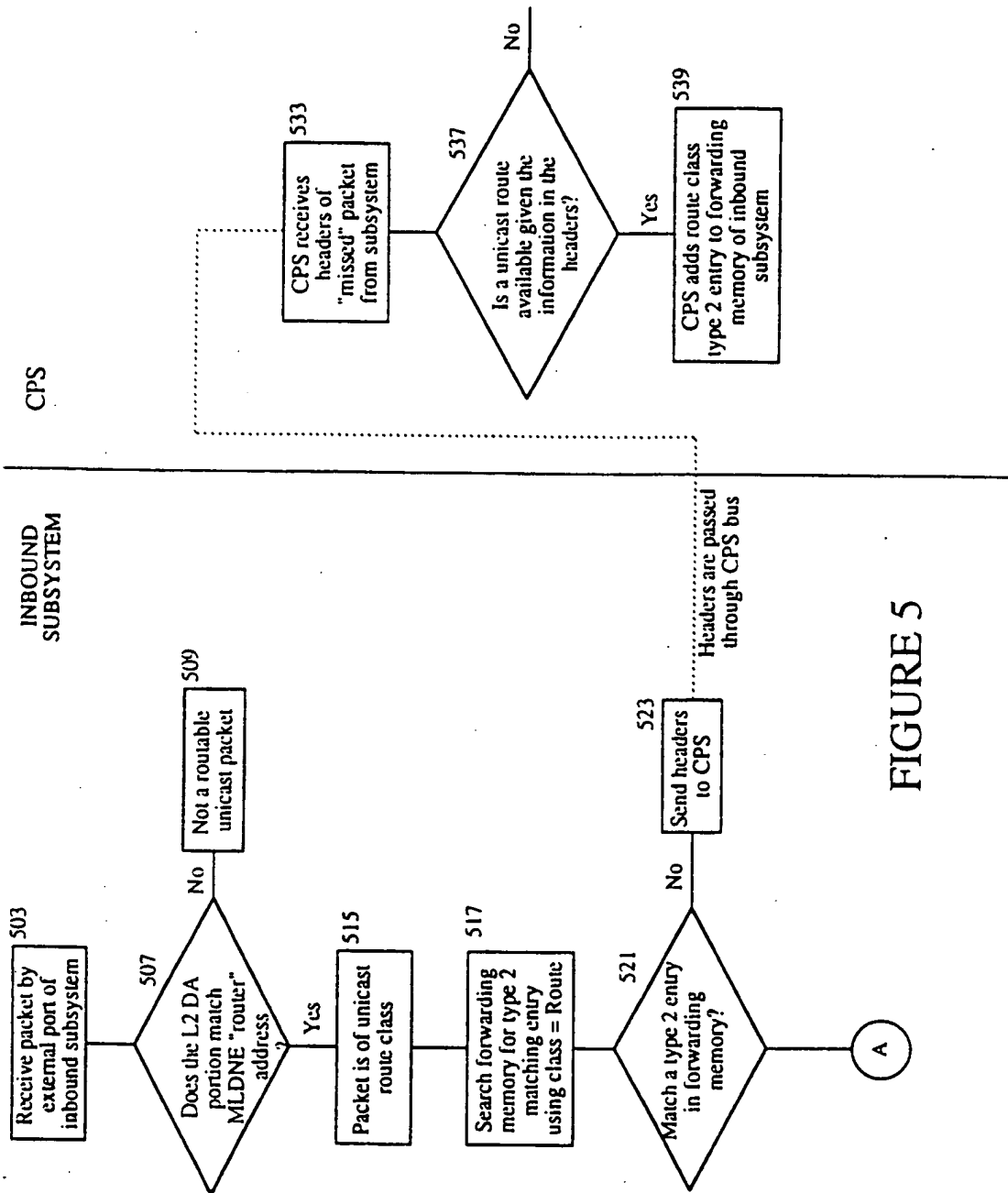
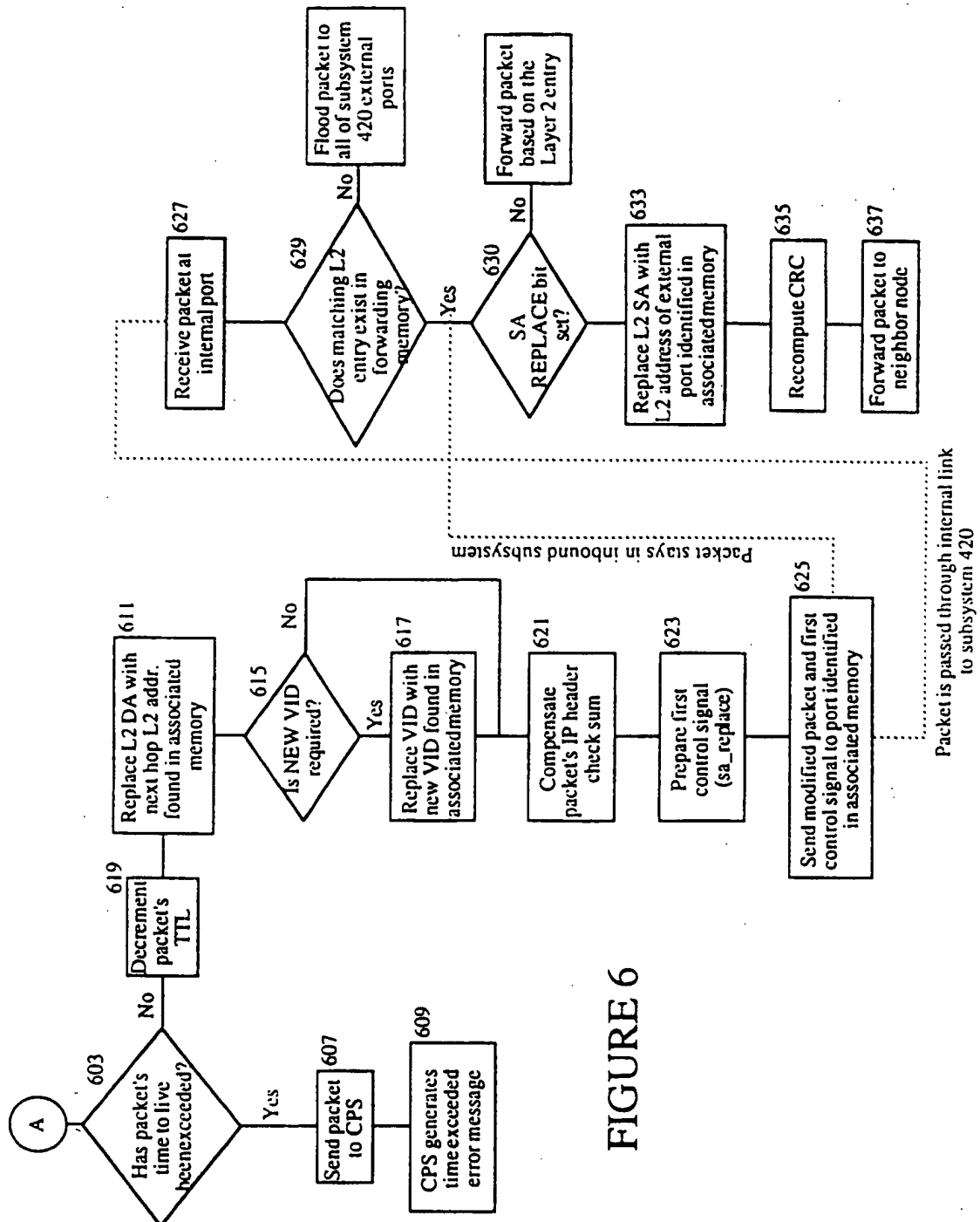


FIGURE 5



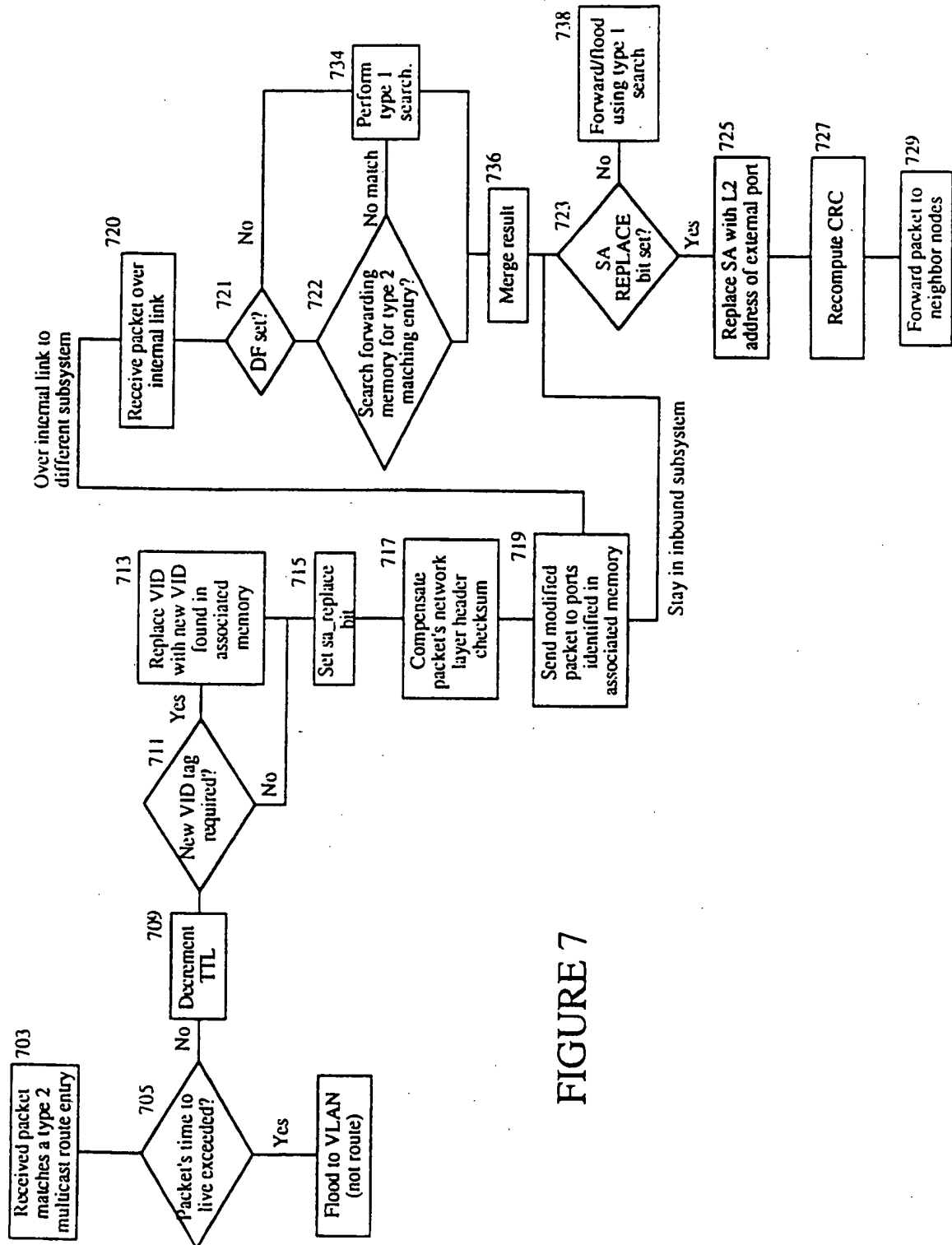


FIGURE 7

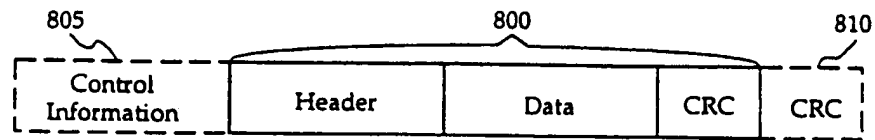


FIGURE 8A

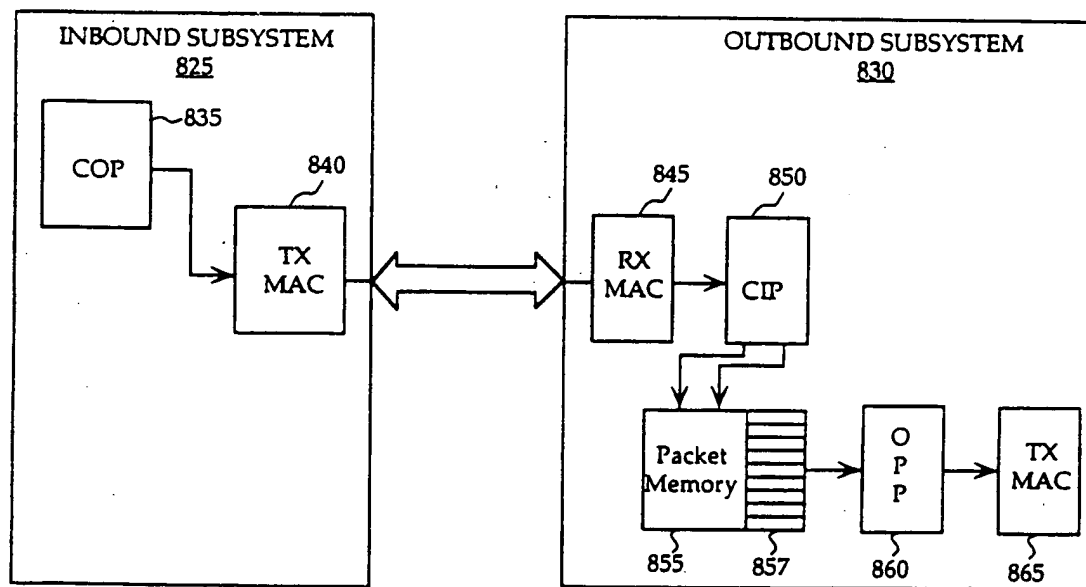


FIGURE 8B

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/13205**A. CLASSIFICATION OF SUBJECT MATTER**

IPC(6) : H04L 12/28, 12/56; H04Q 5/22

US CL : 370/390, 392, 400, 401, 402, 403, 404, 405; 340/825.52

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 370/390, 392, 400, 401, 402, 403, 404, 405; 340/825.52

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,500,860, A (Perlman et al) 19 Mar 1996 (19.03.96), see col. 9, line 22 to col. 12, line 23.	1-18
Y		19-20
Y	US 5,592,476 A (Calamvokis et al.) 7 Jan 1997 (07.01.97), see col. 22, line 34 to col. 23, line 12.	19-20

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:	*T later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
*A document defining the general state of the art which is not considered to be of particular relevance	*X document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
*E earlier document published on or after the international filing date	*Y document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
*L document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	*G document member of the same patent family
*O document referring to an oral disclosure, use, exhibition or other means	
*P document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search

14 SEPTEMBER 1998

Date of mailing of the international search report

19 OCT 1998

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

DUONG, FRANK

Telephone No. (703) 308-5428

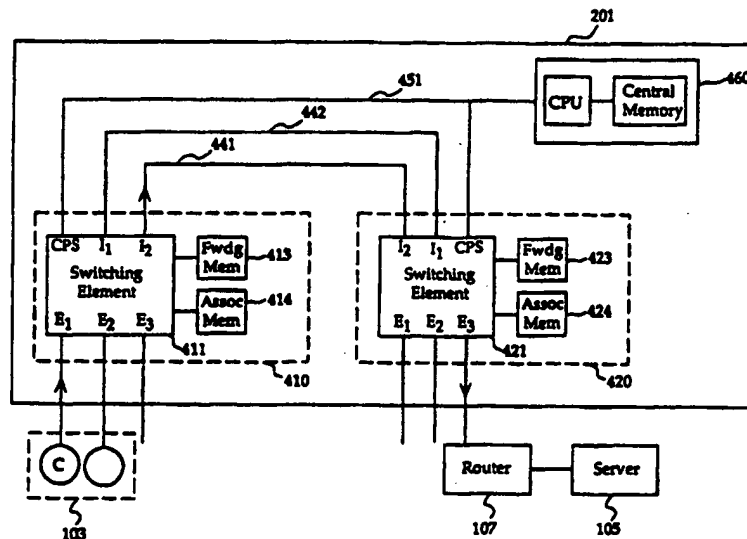
Form PCT/ISA/210 (second sheet)(July 1992)*



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : H04L 12/28, 12/56, H04Q 5/22		A1	(11) International Publication Number: WO 99/00938
			(43) International Publication Date: 7 January 1999 (07.01.99)
(21) International Application Number: PCT/US98/13205		(81) Designated States: JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).	
(22) International Filing Date: 24 June 1998 (24.06.98)		Published <i>With international search report.</i>	
(30) Priority Data: 08/885,114 30 June 1997 (30.06.97) US			
(71) Applicant: SUN MICROSYSTEMS, INC. [US/US]; 901 San Antonio Road, Palo Alto, CA 94303 (US).			
(72) Inventors: HENDEL, Ariel; 7537 Newcastle Drive, Cupertino, CA 95014 (US). MULLER, Shimon; 983 La Mesa Tr., Cupertino, CA 94086 (US). ZAUMEN, William, T.; 912 Clara Drive, Palo Alto, CA 94303 (US). YEUNG, Louise; 110 Rogers Avenue, San Carlos, CA 94070 (US).			
(74) Agents: HYMAN, Eric, S. et al.; Blakely, Sokoloff, Taylor & Zafman, 7th floor, 12400 Wilshire Boulevard, Los Angeles, CA 90025-1026 (US).			

(54) Title: ROUTING IN A MULTI-LAYER DISTRIBUTED NETWORK ELEMENT



(57) Abstract

A multi-layer distributed network element (201) for relaying packets according to known routing protocols. A distributed architecture of multiple subsystems (410) delivers routing at wire-speed performance across subnetworks. Each subsystem (410) includes a forwarding memory (413) and an associated memory (414) and is configured to identify unicast and multicast packets for routing purposes, modify the packets in hardware, including replace VLAN information, and forward the packets to the next hop. The routing decisions are made in the inbound subsystem, and packets are forwarded, if necessary given the network topology, through a separate outbound subsystem.

*(Referred to in PCT Gazette No. 20/1999, Section II)

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						